Risk estimation for matrix recovery with spectral regularization

Charles-Alban Deledalle, Samuel Vaiter, Gabriel Peyré DELEDALLE@CEREMADE.DAUPHINE.FR
CEREMADE, CNRS, Université Paris-Dauphine, France

Jalal Fadili
GREYC, CNRS-ENSICAEN-Université de Caen, France

Charles Dossal
IMB, Université Bordeaux 1, France

Abstract

In this paper, we develop an approach to recursively estimate the quadratic risk for matrix recovery problems regularized with spectral functions. Toward this end, in the spirit of the SURE theory, a key step is to compute the (weak) derivative and divergence of a solution with respect to the observations. As such a solution is not available in closed form, but rather through a proximal splitting algorithm, we propose to recursively compute the divergence from the sequence of iterates. A second challenge that we unlocked is the computation of the (weak) derivative of the proximity operator of a spectral function. To show the potential applicability of our approach, we exemplify it on a matrix completion problem to objectively and automatically select the regularization parameter.

1. Introduction

Consider the problem of estimating a matrix $X_0 \in \mathbb{R}^{n_1 \times n_2}$ from P noisy observations $y = \mathcal{A}(X_0) + w \in \mathbb{R}^P$, where $w \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$. The linear bounded operator $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^P$ entails loss of information such that the problem is ill-posed. This problem arises in various research fields. Because of ill-posedness, side information through a regularizing term is necessary. We thus consider the problem

$$X(y) \in \operatorname*{argmin}_{X \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2}\|y - \mathcal{A}(X)\|^2 + \lambda J(X) \quad (1)$$

where the set of minimizers is non-empty, $\lambda > 0$ is a regularization parameter and $J(X) \in \mathbb{R}$ is a proper lower semi-continuous (lsc) convex regularizing function that imposes the desired structure on $X(y)$. In the following, we focus on the case where J is a convex spectral function, meaning that J depends only on the singular values of its argument. Spectral regularization can account for prior knowledge on the spectrum of X_0, typically low-rank (see e.g. Fazel, 2002).

In practice, the choice of the regularization parameter λ in (1) remains an important problem largely unexplored. Typically, we want to select λ minimizing the quadratic risk $\mathbb{E}_w\|X(y) - X_0\|^2$. Since X_0 is unknown and $X(y)$ is non unique, one can instead consider an unbiased estimate of the *prediction* risk $\mathbb{E}_w\|\mathcal{A}(X(y)) - \mathcal{A}(X_0)\|^2$, where now $\mathcal{A}(X(y))$ is uniquely defined. With the proviso that $\mu(y) = \mathcal{A}(X(y))$ is weakly differentiable, the SURE (for Stein unbiased risk estimator, Stein, 1981)

$$\mathrm{SURE}(y) = \|y - \mu(y))\|^2 - P\sigma^2 + 2\sigma^2 \operatorname{div} \mu(y) \quad (2)$$

is an unbiased estimate of the prediction risk, where $\operatorname{div} \mu(y) = \operatorname{Tr}(\partial \mu(y))$. This unbiased estimate depends solely on y, without prior knowledge of X_0 and then can prove very useful as a basis for automatic ways to choose the regularization parameters λ.

Contributions. Our main contribution is to provide the derivative of matrix-valued spectral functions which extends the result of Lewis & Sendov (2001). This result is used to recursively compute the derivatives of the solutions of spectral regularization from those of the iterates provided by a proximal splitting algorithm. In particular, it provides an estimate of $\operatorname{div} \mu(y)$ in (2) as a basis to compute $\mathrm{SURE}(y)$. A Numerical example on a matrix completion problem is given to support our findings.

2. Recursive risk estimation

Proximal splitting Proximal splitting algorithms have become extremely popular to solve non-smooth convex optimization problems that arise often in inverse problems, e.g. (1). These algorithms provide a sequence of iterates $X^{(\ell)}(y)$ that provably converges to a solution $X(y)$. A practical way to compute $\operatorname{div}\mu(y)$, hence SURE(y), as initiated by Vonesch et al. (2008), and that we pursue here, consists in differentiating this sequence of iterates. This methodology has been extended to a wide class of proximal splitting schemes in (Deledalle et al., 2012b). For the sake of clarity, and without loss of generality, we focus on the case of the forward-backward (FB) splitting algorithm (Combettes & Wajs, 2005).

The FB scheme is a good candidate to solve (1) if J is simple, meaning that its proximity operator has a closed-form. Recall that the proximity operator of a lsc proper convex function G on $\mathbb{R}^{n_1\times n_2}$ is

$$\operatorname{Prox}_G(X) = \operatorname*{argmin}_{Z\in\mathbb{R}^{n_1\times n_2}} \frac{1}{2}\|X-Z\|_F^2 + G(Z).$$

The FB algorithm iteration reads

$$X^{(\ell+1)} = \operatorname{Prox}_{\tau\lambda J}(X^{(\ell)} + \tau\mathcal{A}^*(y-\mathcal{A}(X^{(\ell)}))) \quad (3)$$

where $\mathcal{A}^*$ denotes the adjoint operator of $\mathcal{A}$, $\tau>0$ is chosen such that $\tau\|\mathcal{A}^*\mathcal{A}\|<2$, the dependency of the iterate $X^{(\ell)}$ to y is dropped to lighten the notation.

Risk estimation The divergence term $\operatorname{div}\mu(y)$ is obtained by deriving formula (3), which allows, for any vector $\delta\in\mathbb{R}^P$ to compute iteratively $\xi^{(\ell)} = \partial X^{(\ell)}(y)[\delta]$ (the derivative of $y\mapsto X^{(\ell)}(y)$ at y in the direction δ) as

$$\begin{aligned} \xi^{(\ell+1)} &= \partial\operatorname{Prox}_{\tau\lambda J}(\Xi^{(\ell)})[\zeta^{(\ell)}] \\ \text{where}\quad \Xi^{(\ell)} &= X^{(\ell)} + \tau\mathcal{A}^*(y-\mathcal{A}(X^{(\ell)})) \\ \text{and}\quad \zeta^{(\ell)} &= \xi^{(\ell)} + \tau\mathcal{A}^*(\delta-\mathcal{A}(\xi^{(\ell)})). \end{aligned}$$

Using the Jacobian trace formula of the divergence, it can be easily seen that

$$\operatorname{div}\mu(y) = \mathbb{E}_\delta\langle\partial\mu(y)[\delta],\,\delta\rangle \approx \frac{1}{k}\sum_{i=1}^{k}\langle\partial\mu(y)[\delta_i],\,\delta_i\rangle \quad (4)$$

where $\delta\sim\mathcal{N}(0,\mathrm{Id}_P)$ and δ_i are k realizations of δ. The SURE(y) can in turn be iteratively estimated by plugging $\partial\mu(y)[\delta_i] = \mathcal{A}(\partial X^{(\ell)}(y)[\delta_i])$ in (4).

3. Local behavior of spectral functions

This section studies the local behavior of real- and matrix-valued spectral functions. We write the singular value decomposition (SVD) of a matrix $X\in\mathbb{R}^{n_1\times n_2}$

$$X = V_X\operatorname{diag}(\Lambda_X)U_X^*$$

(which might not be in general unique), where $\Lambda_X\in\mathbb{R}^n$ is the vector of singular values of X with $n=\min(n_1,n_2)$, $\operatorname{diag}(\Lambda_X)\in\mathbb{R}^{n_1\times n_2}$ denotes the matrix with diagonal entries Λ_X, and $V_X\in\mathbb{R}^{n_1\times n_1}$ and $U_X\in\mathbb{R}^{n_2\times n_2}$ are the unitary matrices of singular vectors.

3.1. Real-valued Spectral Function

A spectral function J can by definition be written as

$$J(X) = \varphi(\Lambda_X) \quad (5)$$

where $\varphi:\mathbb{R}^n\to\mathbb{R}$ is a symmetric function of its argument, meaning $\varphi(P\Lambda)=\varphi(\Lambda)$ for any permutation matrix $P\in\mathbb{R}^{n\times n}$ and Λ in the domain of φ. We extend φ to the negative half-line as $\varphi(\Lambda)=\varphi(|\Lambda|)$.

We consider $J:\mathbb{R}^{n_1\times n_2}\to\mathbb{R}$ a real-valued spectral function as defined in (5). From subdifferential calculus on spectral functions (Lewis, 1995), we get the following.

Proposition 1. *A spectral function $J(X)=\varphi(\Lambda_X)$ is convex if and only if φ is convex, and then*

$$\forall\gamma>0,\quad \operatorname{Prox}_{\gamma J}(X) = V_X\operatorname{diag}(\operatorname{Prox}_{\gamma\varphi}(\Lambda_X))U_X^*.$$

3.2. Matrix-valued Spectral Function

We now turn to matrix-valued spectral functions

$$F(X) = V_X\operatorname{diag}(\Phi(\Lambda_X))U_X^*. \quad (6)$$

where $\Phi:\mathbb{R}^n\to\mathbb{R}^n$ is symmetric in its arguments, meaning $\Phi\circ P = P\circ\Phi$ for any permutation matrix $P\in\mathbb{R}^{n\times n}$. We extend Φ to negative numbers as $\Phi(\Lambda)=\operatorname{sign}(\Lambda)\odot\Phi(|\Lambda|)$ and $\odot$ is the entries-wise matrix multiplication. Remark that taking $F(X)=\operatorname{Prox}_{\gamma J}(X)$ and $\Phi=\operatorname{Prox}_{\gamma\varphi}$, the proximity operator of a convex real-valued spectral function is a matrix-valued spectral function. Moreover, the following theorem provides a closed-form expression of the derivative of F that generalizes the result of Lewis & Sendov (2001) to any (non-necessarily symmetric) matrix-valued functions.

Without loss of generality, we consider the case of square matrices, i.e. $n_1=n_2=n$. Substituting X by the square matrix $\tilde{X}\in\mathbb{R}^{m\times m}$ where $m=\max(n_1,n_2)$ and $\tilde{X}_{i,j}=X_{i,j}$ if $i\le n_1$ and $j\le n_2$, 0 otherwise, gives the general result.

Theorem 1. *A squared matrix-valued spectral function F is differentiable at X if and only if Φ is differentiable at Λ_X. In this case*

$$\forall\,\delta\in\mathbb{R}^{n\times n},\quad \partial F(X)[\delta] = V_X H_X[\delta]U_X^*$$
$$\text{with } H_X[\delta] = \mathcal{M}(\delta) + \Gamma_S(\Lambda_X)\odot\mathcal{P}_S(\delta) + \Gamma_A(\Lambda_X)\odot\mathcal{P}_A(\delta)$$

where the symmetric and anti-symmetric parts are

$$\mathcal{P}_S(Y) = \frac{Y + Y^*}{2} \quad \textit{and} \quad \mathcal{P}_A(Y) = \frac{Y - Y^*}{2}$$

and $\mathcal{M} : \mathbb{R}^n \mapsto \mathbb{R}^{n\times n}$ *is*

$$\mathcal{M} = \operatorname{diag} \circ \partial\Phi(\Lambda_X) \circ \operatorname{diag}.$$

The matrices $\Gamma_S(\Lambda)$ *and* $\Gamma_A(\Lambda)$ *are defined as*

$$\Gamma_S(\Lambda)_{i,j} = \begin{cases} 0 & \textit{if} \quad i = j \\ \frac{\Phi(\Lambda)_i - \Phi(\Lambda)_j}{\Lambda_i - \Lambda_j} & \textit{if} \quad \Lambda_i \neq \Lambda_j \\ \partial\Phi(\Lambda)_{i,i} - \partial\Phi(\Lambda)_{i,j} & \textit{otherwise,} \end{cases}$$

$$\Gamma_A(\Lambda)_{i,j} = \begin{cases} 0 & \textit{if} \quad i = j \\ \frac{\Phi(\Lambda)_i + \Phi(\Lambda)_j}{\Lambda_i + \Lambda_j} & \textit{otherwise.} \end{cases}$$

A proof summary is given in Appendix A. Note that for symmetric matrices X and δ, we recover the result of Lewis & Sendov (2001).

4. Numerical applications

4.1. Nuclear norm regularization

We here consider the problem of recovering a low-rank matrix $X_0 \in \mathbb{R}^{n_1\times n_2}$. To this end, J is taken as the nuclear norm (a.k.a., trace or Schatten 1-norm) which is in some sense the tightest convex relaxation to the NP-hard rank minimization problem (Candès & Recht, 2009). The nuclear norm is defined by

$$J(X) = \|X\|_* \triangleq \|\Lambda_X\|_1. \tag{7}$$

Taking $J(\cdot)$ as $\|\cdot\|_*$ and φ as $\|.\|_1$ in Proposition 1 gives:

Corollary 1. *The proximal operator of* $\gamma\|\cdot\|_*$ *is*

$$\forall \gamma > 0, \quad \operatorname{Prox}_{\gamma\|\cdot\|_*}(X) = V_X \operatorname{diag}(T_\gamma(\Lambda_X)) U_X^* \ , \tag{8}$$

where $T_\gamma = \operatorname{Prox}_{\gamma\|.\|_1}$ *is the component-wise soft-thresholding, defined for* $i = 1, \ldots, n$ *as*

$$T_\gamma(t)_i = \max(0, 1 - \gamma/\|t_i\|)t_i.$$

We now turn to the derivative of $F = \operatorname{Prox}_{\gamma\|\cdot\|_*}$. A straightforward attempt is to take $\Phi = \operatorname{Prox}_{\gamma\|.\|_1} = T_\gamma$ and apply Theorem 1 with

$$\partial\Phi(X)[\delta]_i = \partial T_\gamma(t)[\delta]_i = \begin{cases} 0 & \text{if} \quad \|t_i\| \leqslant \gamma \\ \delta_i & \text{otherwise.} \end{cases} \tag{9}$$

However, strictly speaking, Theorem 1 does not apply since a proximity mapping is 1-Lipschitz in general, hence not necessarily differentiable everywhere. Thus, its derivative may be set-valued, as is the case for soft-thresholding at $\pm\gamma$.

A direct consequence of Corollary 1 is that J is a simple function allowing for the use of the FB algorithm. Moreover, the expression of the derivative (9) provides an estimation of the SURE as explained in Section2.

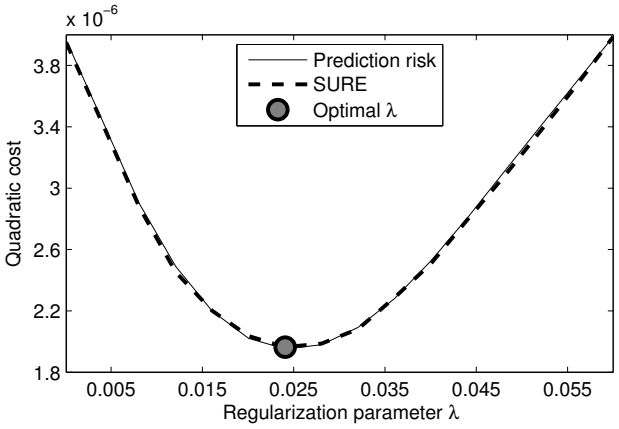

Figure 1. Predicted risk and its SURE estimate[1].

4.2. Application to matrix completion

We now exemplify the proposed SURE computation approach on a matrix completion problem encountered in recommendation systems such as the popular Netflix problem. We therefore consider the forward model $y = \mathcal{A}(X_0) + w \in \mathbb{R}^P$, $w \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_P)$, where X_0 is a dense but low-rank (or approximately so) matrix and A is binary masking operator.

We have taken $(n_1, n_2) = (1000, 100)$ and $P = 25000$ observed entries (i.e., 25%). The underlying dense matrix X_0 has been chosen to be approximately low-rank with a rapidly decaying spectrum $\Lambda_{X_0} = \{k^{-1}\}_{k=1}^n$. The standard deviation σ has been set such that the resulting minimum least-square estimate has a relative error $\|X_{LS} - X_0\|_F / \|X_0\|_F = 0.9$. Figure 1 depicts the prediction risk and its SURE estimate as a function of λ. For each value of λ in the tested range, SURE(y) in (2) has been computed for a single realization of y with $k = 4$ realizations δ_i in (4)[1]. At the optimal λ value, $X(y)$ has a rank of 55 with a relative error of 0.46 (i.e., a gain of about a factor 2 w.r.t. the least-square estimator).

5. Conclusion

The core theoretical contribution of this paper is the derivative of (rectangular) matrix-valued spectral functions. This was a key step to compute the derivative of the proximal operator associated to the nuclear norm, and finally to use the SURE to recursively estimate the quadratic prediction risk of matrix recovery problems involving the nuclear norm regularization. The SURE was also used to automatically select the optimal regularization parameter.

[1]Without impacting the optimal choice of λ, the two curves have been vertically shifted for visualization.

A. Summary of the proof of Theorem 1

We prove Theorem 1 in the case of X with distinct singular values. Due to obvious lack of space, the proof for the case of multiple singular values being more difficult is achieved in (Deledalle et al., 2012a).

The following lemma derives the expression of the derivative of the SVD mapping $X \mapsto (V_X, \Lambda_X, U_X)$. Note that this mapping is not well defined because even if the Λ_X are distinct, one can apply arbitrary sign changes and permutations to the set of singular vectors. The lemma should thus be interpreted in the sense that one can locally write a Taylor expansion using the given differential for any particular choice of SVD. For a more elegant statement of this lemma using wedge products, one can refer to (Edelman, 2005).

Lemma 1. *We consider X with distinct singular values. The singular value mapping admit a local Taylor expansion at X. Then for a given matrix δ, the derivative of the singular value mapping $x \mapsto \Lambda_X$ is*

$$\partial\Lambda_X[\delta] = \operatorname{diag}(V_X^*\delta U_X). \tag{10}$$

To state the derivative of the singular vector mapping, we write $\delta_V = V_X^*\partial V_X[\delta]$ and $\delta_U = U_X^*\partial U_X[\delta]$, and also $\Lambda = \Lambda_X$, ignoring the dependency with respect to X. Then

$$(\delta_V)_{i,j} = \frac{\Lambda_j\bar{\delta}_{i,j} + \Lambda_i\bar{\delta}_{j,i}}{\Lambda_j^2 - \Lambda_i^2} \quad \text{and} \quad (\delta_U)_{i,j} = \frac{\Lambda_i\bar{\delta}_{i,j} + \Lambda_j\bar{\delta}_{j,i}}{\Lambda_j^2 - \Lambda_i^2} \tag{11}$$

where for some Y we denote $\bar{Y} = V_X^*YU_X$.

Proof. Deriving the relationship $X = V_X \operatorname{diag}(\Lambda_X)U_X^*$ and multiplying on the left by V_X^* and on the right by U_X gives

$$\bar{\delta} = \delta_V \operatorname{diag}(\Lambda_X) + \operatorname{diag}(\Lambda_X)\delta_U^* + \partial\operatorname{diag}(\Lambda_X)[\delta]. \tag{12}$$

The relation $U_XU_X^* = \mathrm{Id}$ and $V_XV_X^* = \mathrm{Id}$ implies that the matrices δ_U and δ_V are antisymmetric. In particular they are zero along the diagonal. Thus applying the operator diag to both sides of (12) shows (10). Now considering the entries (i,j) and (j,i) of the linear system (12) shows that one needs to solve a series of 2×2 symmetric linear systems

$$\begin{pmatrix} \Lambda_j & -\Lambda_i \\ -\Lambda_i & \Lambda_j \end{pmatrix} \begin{pmatrix} (\delta_V)_{i,j} \\ (\delta_U)_{i,j} \end{pmatrix} = \begin{pmatrix} \bar{\delta}_{i,j} \\ \bar{\delta}_{j,i} \end{pmatrix}.$$

This system can be solved in closed form, which gives the desired formula □

We now prove the theorem in the case of X with distinct singular values.

Proof. Assume that the singular values of X are all distinct. Differentiating the relationship (6) gives

$$V_X^*\partial F(X)[\delta]U_X = \delta_V \operatorname{diag}(\Phi(\Lambda_X)) + \operatorname{diag}(\Phi(\Lambda_X))\delta_U^* + \mathcal{M}(\delta_X)$$

where we have used the notation introduced in Lemma 1. Using the expression (11) for δ_U and δ_V shows that the matrix $W = \delta_V \operatorname{diag}(\Phi(\Lambda_X)) + \operatorname{diag}(\Phi(\Lambda_X))\delta_U^*$ is computed as

$$W_{i,j} = \frac{\varphi_j(\Lambda_j\bar{\delta}_{i,j} + \Lambda_i\bar{\delta}_{j,i}) - \varphi_i(\Lambda_i\bar{\delta}_{i,j} + \Lambda_j\bar{\delta}_{j,i})}{\Lambda_j^2 - \Lambda_i^2}$$

where $\varphi = \Phi(\Lambda)$. Rearranging this expression using the symmetric and anti-symmetric parts shows the desired formula when F has distinct singular values. □

References

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

Combettes, P. L. and Wajs, V. R. Signal recovery by proximal forward-backward splitting. *Math. Mod. Sim.*, 4(4): 1168, 2005.

Deledalle, C., Peyré, G., and Mirebeau, J.M. Derivatives of matrix-valued spectral functions. Technical report, 2012a.

Deledalle, C., Vaiter, S., Peyré, G., Fadili, J., and Dossal, C. Proximal Splitting Derivatives for Risk Estimation. Technical report, Feb. 2012b. URL http://hal.archives-ouvertes.fr/hal-00670213.

Edelman, A. Matrix jacobians with wedge products. *MIT Handout for 18.325*, 2005.

Fazel, M. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

Lewis, A.S. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1/2):173–183, 1995.

Lewis, A.S. and Sendov, H.S. Twice differentiable spectral functions. *SIAM Journal on Matrix Analysis on Matrix Analysis and Applications*, 23:368–386, 2001.

Stein, C.M. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Vonesch, C., Ramani, S., and Unser, M. Recursive risk estimation for non-linear image deconvolution with a wavelet-domain sparsity constraint. In *ICIP*, pp. 665–668. IEEE, 2008.